Identifying Morality Frames in Political Tweets using Relational Learning

Shamik Roy, Maria Leonor Pacheco, Dan Goldwasser
Department of Computer Science
Purdue University
West Lafayette, IN, USA
{roy98, pachecog, dgoldwas}@purdue.edu

Abstract

Extracting moral sentiment from text is a vital component in understanding public opinion, social movements, and policy decisions. The Moral Foundation Theory identifies five moral foundations, each associated with a positive and negative polarity. However, moral sentiment is often motivated by its targets, which can correspond to individuals or collective entities. In this paper, we introduce morality frames, a representation framework for organizing moral attitudes directed at different entities, and come up with a novel and high-quality annotated dataset of tweets written by US politicians. Then, we propose a relational learning model to predict moral attitudes towards entities and moral foundations jointly. We do qualitative and quantitative evaluations, showing that moral sentiment towards entities differs highly across political ideologies.

1 Introduction

Morality is a set of principles to distinguish between right and wrong. Shared moral values form the social and cultural norms that unite social groups (Dehghani et al., 2016). *Moral Foundations Theory* (MFT) (Haidt and Joseph, 2004; Haidt and Graham, 2007) provides a theoretical framework for analyzing different expressions of moral values. The theory suggests that there are at least five basic moral values, emerging from evolutionary, social, and cultural origins. These are referred to as Moral Foundations (MFs), each with a positive and a negative polarity, and include *Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/-Subversion*, and *Purity/Degradation* (Table 1 provides details). Identifying MF in text is a relatively new challenge and past work has relied on lexical resources such as the Moral Foundation Dictionary (Graham et al., 2009; Fulgoni et al., 2016; Xie et al., 2019) and annotated data (Johnson and Goldwasser, 2018; Lin et al., 2018; Hoover et al., 2020).

Social and political science studies have repeatedly shown the correlation between ideological and political stances and moral foundation preferences (Graham et al., 2009; Wolsko et al., 2016; Amin et al., 2017). For example, Graham et al., 2009 captures the correlation between political ideology and moral foundation usage, showing that Liberals have a preference for Care/Harm and Fairness/Cheating while Conservatives use all five. Our main intuition in this paper is that even when different groups use the same MF, the moral sentiment would be directed towards different targets. To clarify, consider the following tweets discussing the Affordable Healthcare Act (ACA, Obamacare).

[@SenThadCochran and I]$_{CARING}$ are working to protect [MS small businesses]$_{CARE-FOR}$ from more expensive [#Obamacare mandates]$_{HARMING}$.

[The ACA]$_{CARING}$ was a life saver for the more than [130 million Americans]$_{CARE-FOR}$ with a preexisting condition – including covid now. [Republicans]$_{HARMING}$ want to take us back to coverage denials.

While both tweets use the Care/Harm MF, in the top tweet (Conservative) the ACA is described as causing Harm, while in the bottom (Liberal), the ACA is described as providing the needed Care.

Our main contribution in this paper is to introduce *morality frames*, a representation framework for organizing moral attitudes directed at different targets, by decomposing the moral foundations into structured frames, each associated with a predicate (a specific MF) and typed roles. For example, the morality frame for Care/Harm is associated with three typed roles: entity providing care, entity needing the care, and entity causing harm. We focus on analyzing political tweets, each describing an eliciting situation that evokes the moral sentiment, and map the text to a MF, and the entities appearing in it to typed roles. Given tweets by different ideological groups discussing the same real-world situation, morality frames can provide the means to explain and compare the attitudes of the two groups.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9939–9958
November 7–11, 2021.

We build on the MF dataset by Johnson and Goldwasser, 2018 consisting of political tweets, and annotate each tweet with MF roles for its entities.

Identifying moral roles from text in our setting requires inferences based on *political knowledge*, mapping between the author's perspectives and the judgements appearing in the text. For example, Donald Trump is likely to elicit a negative moral judgement from most Liberals and a positive one from most Conservatives, regardless of the specific moral foundation that is evoked. From a technical perspective, our goal is to model these kind of dependencies in a probabilistic framework, connecting MF and roles assignments, entity-specific sentiment polarity and repeating patterns within ideological groups (while our focus is U.S. politics, these settings could be easily adapted to capture patterns based on other criteria). We formulate these dependencies as a structured learning task and compare two relational learning frameworks, PSL (Bach et al., 2017) and DRaiL (Pacheco and Goldwasser, 2021). Our experiments demonstrate that modeling these dependencies, capturing political and social knowledge, result in improved performance. In addition, we conduct a thorough ablation study and error analysis to explain their impact on performance.

Finally, we demonstrate how entity-based MF analysis can help capture perspective differences based on ideological lines. We apply our model to tweets by members of Congress on the issue of Abortion and the 2021 storming of the US Capitol. Our analysis shows that while Conservative and Liberal tweets target the same entities, their attitudes are often conflicting.

2 Related Work

Usage of sociological theories like the Moral Foundation Theory (MFT) (Haidt and Joseph, 2004; Haidt and Graham, 2007) and Framing (Entman, 1993; Chong and Druckman, 2007; Boydstun et al., 2014) in Natural Language Processing tasks has gained significant interest. The Moral Foundation Theory (MFT) has been widely used to study the effect of moral values on human behavioral patterns, such as charitable donations (Hoover et al., 2018), violent protests (Mooijman et al., 2018) and social homophily (Dehghani et al., 2016). Framing is a strategy used to bias the discussion on an issue towards a specific stance by emphasizing certain aspects that prime the reader to support the stance. Framing is used to study the political bias and polarization in social and news media (Tsur et al., 2015; Baumer et al., 2015; Card et al., 2015; Field et al., 2018; Demszky et al., 2019; Fan et al., 2019; Roy and Goldwasser, 2020). Moral Foundation Theory (MFT) is frequently used to analyze political framing and agenda setting. For example, Fulgoni et al. (2016) analyzed framing in partisan news sources using MFT, Dehghani et al. (2014) studied the difference in moral sentiment usage between liberals and conservatives. Brady et al. (2017) found that moral/emotional political messages are diffused at higher rates on social media.

Previous works have also contributed to the detection of moral sentiments. Johnson and Goldwasser (2018) showed that policy frames (Boydstun et al., 2014) help in moral foundation prediction, Hoover et al. (2020) proposed a dataset of 35k tweets annotated for moral foundations, Lin et al. (2018) used background knowledge for moral sentiment prediction, Xie et al. (2019) proposed a text based framework to account for moral sentiment change, and Garten et al. (2016) used pretrained distributed representations of words to extend the Moral Foundations Dictionary (Graham et al., 2009) for detecting moral rhetoric.

While existing works study MFT at the issue and sentence level, Roy and Goldwasser (2021) showed that there is a correlation between entity mention and the sentence-level moral foundation in the tweets by the U.S. politicians. We extend this work by studying MFT directly at the entity level. Hence, our work is broadly related to the works on entity-centric affect analysis (Deng and Wiebe, 2015; Field and Tsvetkov, 2019; Park et al., 2020).

Combining neural networks and structured inference was explored for traditional NLP tasks such as dependency parsing (Chen and Manning, 2014; Weiss et al., 2015; Andor et al., 2016), named entity recognition (Lample et al., 2016) and sequence labeling systems (Ma and Hovy, 2016; Zhang et al., 2017). Recently, these efforts have expanded to discourse-level tasks such as argumentation mining (Niculae et al., 2017; Widmoser et al., 2021), event/temporal relation extraction (Han et al., 2019) and discourse representation parsing (Liu et al., 2019). Following this trend, Pacheco and Goldwasser (2021) introduced DRaiL, a general declarative framework for deep structured prediction, designed specifically for NLP tasks. In this paper, we use DRaiL to model moral foundations

and morally-targeted entities in tweets, and find an improvement over other non-neural probabilistic graphical modeling frameworks (Bach et al., 2017).

MORAL FOUNDATIONS	MORAL ROLES
CARE/HARM: Care for others, generosity, compassion, ability to feel pain of others, sensitivity to suffering of others, prohibiting actions that harm others.	1. Target of care/harm 2. Entity causing harm 3. Entity providing care
FAIRNESS/CHEATING: Fairness, justice, reciprocity, reciprocal altruism, rights, autonomy, equality, proportionality, prohibiting cheating.	1. Target of fairness/cheating 2. Entity ensuring fairness 3. Entity doing cheating
LOYALTY/BETRAYAL: Group affiliation and solidarity, virtues of patriotism, self-sacrifice for the group, prohibiting betrayal of one's group.	1. Target of loyalty/betrayal 2. Entity being loyal 3. Entity doing betrayal
AUTHORITY/SUBVERSION: Fulfilling social roles, submitting to authority, respect for social hierarchy/traditions, leadership, prohibiting rebellion against authority.	1. Justified authority 2. Justified authority over 3. Failing authority 4. Failing authority over
PURITY/DEGRADATION: Associations with the sacred and holy, disgust, contamination, religious notions which guide how to live, prohibiting violating the sacred.	1. Target of purity/degradation 2. Entity preserving purity 3. Entity causing degradation

Table 1: Moral foundations and their associated roles.

3 Identifying Entity-Centric Moral Roles

3.1 Morality Frames

MFT defines a convenient abstraction of the moral sentiment expressed in a given text. **Morality Frames** build on MFT and provide *entity-centric moral sentiment* information. Rather than defining negative and positive MF polarities (e.g., CARE or HARM), we use the five MFs as frame predicates, and associate positive and negative entity roles with each frame. As described in Table 1, these roles capture information specific to each MF. For example, *entity causing harm*, is a negative sentiment role, associated with the CARE/HARM morality frame. The entities filling these roles can be individuals, collective entities, objects, activities, concepts, or legislative elements.

3.2 Data Collection

We build on the dataset proposed by Johnson and Goldwasser (2018), consisting of tweets by US politicians posted between 2016 and 2017. A subset of it (2K out of 93K) is annotated for Moral Foundations and Policy Frames (Boydstun et al., 2014). The tweets focus on six politically polarized issues: *immigration, guns, abortion, ACA, LGBTQ, and terrorism*, and the party affiliations of the authors are known. We consider only labeled moral tweets, and choose the most prominent MF annotation for each tweet (some tweets are annotated for a secondary MF). Since the data contains only few examples of the *Purity/Degradation* moral foundation, we collected more examples from the unlabeled segment and manually annotated them. Table 2 shows the statistics of the final dataset. The annotation process and per-topic distribution of tweets are outlined in Appendix A.

MORALS	# OF TWEETS	IDEOLOGY	
		LEFT	RIGHT
Care/Harm	589	378	211
Fairness/Cheating	264	201	63
Loyalty/Betrayal	231	167	64
Authority/Subversion	471	200	271
Purity/Degradation	44	13	31
TOTAL	1599	959	640

Table 2: Dataset summary.

3.3 Entity Roles Annotation

We annotate each tweet for entities and their associated moral roles.

Annotation Schema: We set up a QA task on Amazon Mechanical Turk. Annotators were given a tweet, the associated MF label and its description. They were then presented with multiple questions, and asked to mark the answers, corresponding to our entity roles, in the tweet. Table 3 shows the questions asked for the *Care/Harm* case. We asked additional questions to assess the annotators' understanding of the task. The questions for other moral foundations can be found in Appendix B.1.

Quality Assurance: We provided the annotators with work-through examples and hints with each question about the entity type. The interface allowed them to mark a segment of the text with one moral role only. To further improve the quality, we did the annotation in two phases. In the annotator selection phase, we released a small subset of tweets for annotation. Based on the annotations, we assigned qualifications to high performing workers and released the rest of the tweets only to them. We awarded the annotators $15 - 18¢$ per tweet. We define agreement among annotators if they mark the same segment in the text as having the same entity-role. We calculate the agreement among multiple annotators using Krippendorff's α (Krippendorff, 2004), where $\alpha = 1$ means perfect agreement,

Entity Type	Question Asked to the Annotators
Target of care/harm	Which entity needs care, or is being harmed?
Entity causing harm	Which entity is causing the harm?
Entity provid. care	Which entity is offering/providing the care?

Table 3: Questionnaire for entity roles in care/harm.

MORALS	# TWEETS		# ANN/TW		AGREEMENT (SD)	
	S	F	S	F	S	F
Care/Harm	27	589	3	3	0.63 (0.5)	0.70 (0.5)
Fairness/Cheating	30	247	5.03	2.92	0.55 (0.4)	0.69 (0.5)
Loyalty/Betrayal	40	203	5.67	2.89	0.58 (0.3)	0.63 (0.5)
Authority/Subversion	50	466	4.58	2.92	0.55 (0.3)	0.60 (0.5)
Purity/Degradation	10	36	6	3	0.51 (0.4)	0.77 (0.6)

Table 4: Annotator agreement in selection (S) and final (F) phases. ANN/TW refers to number of annotations per tweet and SD refers to Standard Deviation.

$\alpha = -1$ means inverse agreement, and $\alpha = 0$ is the level of chance in a tweet. Table 4 shows that the average agreement increased in the final stage. Note that the annotator agreement (Krippendorff's α) is calculated by comparing the character by character agreement between annotations. For example, if one annotator has marked 'President Trump' as an answer in a tweet, and another has marked 'Trump' as the answer, it will be considered as agreement on the characters 'Trump' but disagreement on 'President', although they really did not disagree on their annotations. This makes the agreement measurement very strict. Regardless, we still got very good average agreement among annotators in the final annotation step. We further refine the annotations by taking majority voting as described in the following section.

Annotation Results: A tweet is annotated by at least three annotators. We define a text span to be an entity E, having a moral role M, in tweet T, if it is annotated as such by at least two annotators. This way, we found $2,945$ (T, E, M) tuples.

To compare the partisanship of MFs and MF roles, we calculate the z-scores for the proportion of MFs and MF roles in the left and right, and consider it as partisan score (- right, + left). The partisan scores for common MFs and their corresponding most partisan (role: entity) tuples are shown in Table 5. The results of this analysis align with our intuition, moral sentiment towards entities can be more indicative of partisanship than the high-level MFs. In Table 6, we present the top-5 most used entities per role by political party for *Care/Harm*. We can see that the target entities of moral roles vary significantly across parties. Details for other MFs and z-scores are in Appx. B.2.

Topic	Common MF (Partisan score)	Most Partisan Role:Entity	
		Right (-)	Left (+)
Abort	Care/Harm (+1.4)	Harming: PPFA (-3.2)	Caring: PPFA (+0.5)
Immig	Auth/Subv (-6.8)	Failing Authority: Obama (-0.5)	Failing Authority: SCOTUS (+5.3)
Guns	Care/Harm (+0.6)	Care For: Law Abiding Citizens (-4.8)	Harming: Gun (+4.1)
ACA	Care/Harm (+2.2)	Harming: ACA (-5.7)	Caring: ACA (+5.6)
Terror	Care/Harm (+2.0)	Harming: Islam (-2.4)	Harming: Terr. Suspect (+3.5)
LGBT	Fair/Cheat (+2.1)	Cheating: SCOTUS-Marriage (-6.5)	Target of Fairness: LGBT (+1.0)

Table 5: Partisanship of MF and MF Roles.

Entity Types	Most Frequent Entities in Left	Most Frequent Entities in Right
Target of care/harm	20 million Americans; our families; woman; innocent people; #domesticviolence victims	law-abiding Americans; victims and their families; small businesses; patients; Paris
Entity causing harm	gun show loopholes; gun violence; terrorist attack; mass-shootings; suspected terrorists	Radical Islamic terrorists; #Obamacare mandates; Brussels attacks; #ISIS; ISIL-Inspired Attacks
Entity providing care	gun show loophole bills; Affordable Care Act; #ImmigrationReform; Democrats; commonsense gun legislation	@RepHalRogers: Bill; @HouseGOP; Senate; @WaysandMeansGOP; HR 240

Table 6: Top-5 frequent entities by role for Care/Harm.

4 Model

We propose a relational learning model for identifying morality frames in text. We begin by defining our relational structure (Sec. 4.1) and proceed to describe its implementation using relational learning tools (Sec. 4.2).

4.1 Relational Model for Morality Frames

Statistical Relational Learning (SRL) methods attempt to model a joint distribution over relational data, and have proven useful in tasks where contextualizing information and interdependent decisions can compensate for a low number of annotated examples (Deng and Wiebe, 2015; Johnson and Goldwasser, 2016, 2018; Subramanian et al., 2018). By breaking down the problem into interdependent relations, these approaches are easier to interpret than end-to-end deep learning techniques.

We propose a joint prediction framework of morality frames, modeling the dependency between MF labels and moral roles instances. Following SRL conventions (Richardson and Domingos, 2006; Bach et al., 2017), we use first-order-logic to describe relational properties. Specifically, a logical rule is used to define a probabilistic scoring

function over the relation instances appearing in it, the full description appears in Section 4.2.

$$r_1 : \texttt{Tweet}(\texttt{t}) \Rightarrow \texttt{MF}(\texttt{t}, \texttt{m})$$
$$r_2 : \texttt{Tweet}(\texttt{t}) \wedge \texttt{Ent}(\texttt{t}, \texttt{e}) \Rightarrow \texttt{Role}(\texttt{t}, \texttt{e}, \texttt{r})$$

In addition, we make the observation that both moral foundations and entities' moral roles depend on external factors that go beyond the text, such as author information and party affiliation. Previous work has shown that explicitly modeling party affiliation and the topics discussed are helpful for predicting moral foundations (Johnson and Goldwasser, 2018). For this reason, we condition both the moral foundation and moral roles on this additional information, as shown in the rules below.

$$r_3 : \texttt{Tweet}(\texttt{t}) \wedge \texttt{Ideo}(\texttt{t}, \texttt{i}) \wedge \texttt{Topic}(\texttt{t}, \texttt{k}) \Rightarrow \texttt{MF}(\texttt{t}, \texttt{m})$$
$$r_4 : \texttt{Tweet}(\texttt{t}) \wedge \texttt{Ideo}(\texttt{t}, \texttt{i}) \wedge \texttt{Topic}(\texttt{t}, \texttt{k}) \wedge \texttt{Ent}(\texttt{t}, \texttt{e}) \Rightarrow \texttt{Role}(\texttt{t}, \texttt{e}, \texttt{r})$$

Rules r_1, r_2 condition the moral foundation label (m) and moral foundation role label (r) on the tweet (t) and entity (e), while r_3, r_4 condition on the ideology of the author (i) and the topic of the tweets (k). Concretely, r_4 can be translated as *"if a tweet* t *has author ideology* i*, topic* k*, and mentions entity* e*, the entity will have moral role* r*"*. Other rules can be translated similarly. Then, we explicitly model the dependencies among different decisions using the following three constraints.

$$c_1 : \texttt{Ent}(\texttt{t}, \texttt{e}) \wedge \texttt{Role}(\texttt{e}, \texttt{r}) \wedge \texttt{MF\_Role}(\texttt{m}, \texttt{r}) \Rightarrow \texttt{MF}(\texttt{t}, \texttt{m})$$
$$c_2 : \texttt{Ent}(\texttt{t}, \texttt{e}_1) \wedge \texttt{Ent}(\texttt{t}, \texttt{e}_2) \wedge \texttt{Role}(\texttt{t}, \texttt{e}_1, \texttt{r}) \Rightarrow \neg\texttt{Role}(\texttt{t}, \texttt{e}_2, \texttt{r})$$
$$c_3 : \texttt{SameIdeo}(\texttt{t}_1, \texttt{t}_2) \wedge \texttt{SameTopic}(\texttt{t}_1, \texttt{t}_2) \wedge \texttt{Ent}(\texttt{t}_1, \texttt{e}) \wedge \texttt{Ent}(\texttt{t}_2, \texttt{e}) \wedge \texttt{Role}(\texttt{t}_1, \texttt{e}, \texttt{r}_1) \wedge \texttt{Role}(\texttt{t}_2, \texttt{e}, \texttt{r}_2) \Rightarrow \texttt{SamePolarity}(\texttt{r}_1, \texttt{r}_2)$$

(c_1) **Consistency between MF label and roles:** While rules r_1, r_3 predict the MF labels, and r_2, r_4 predict the role labels, these two predictions are interdependent. Knowing the MF of a tweet limits the space of feasible roles. Likewise, knowing the role of an entity in a tweet will directly give us the MF label. For example, the presence of an entity frequently used as a harming entity indicates a higher probability of the MF label 'Care/Harm'. We model the dependency between these two decisions using constraint c_1, which can be translated as *"if an entity* e*, mentioned in tweet* t*, has role* r*, tied to MF* m*, then tweet* t *will have MF label* m*"*.

(c_2) **Different roles for different entities in the same tweet:** Our intuition is that if multiple entities are mentioned in the same tweet, they are likely to have different roles. While this may not always hold true, we use this constraint to prevent the model from relying only on textual context, and assigning the same role to all entities.

(c_3) **Consistency in the polarity of sentiment towards an entity within a political party:** Intuitively, role types have a positive or negative sentiment associated to them. For example, an entity causing harm and an entity doing betrayal carry negative sentiment. Intuitive polarity for each MF role can be found in Appendix C.1. Given the highly polarized domain that we are dealing with, we assume that regardless of the MF, an entity will likely maintain the same polarity when mentioned by a specific political party across the same topic. Constraint c_3 encourages this consistency, and it can be translated as: *"if two tweets* $\texttt{t}_1$, $\texttt{t}_2$ *are written by authors of the same political ideology, on the same topic, and mention the same entity* e*, then the polarity of the roles* $\texttt{r}_1$ *and* $\texttt{r}_2$ *of* e *in both tweets will likely be the same."* We consider two entities to be the same if they are an exact lexical match, and leave entity clustering for future work.

4.2 Frameworks for Relational Learning

In this work, we experiment with two existing frameworks for modeling relational learning problems - (1) Probabilistic Soft Logic (PSL) (Bach et al., 2017) and (2) Deep Relational Learning (DRaiL) (Pacheco and Goldwasser, 2021). Both PSL and DRaiL are probabilistic frameworks for specifying and learning relational models using weighted logical rules, specifically horn clauses of the form $w_r : P_1 \wedge ... \wedge P_{n-1} \Rightarrow P_n$. Weights w_r indicate the importance of each rule in the model, and they can be learned from data. Predicates P_i can be closed, if they are observed, or open if they are unobserved. Probabilistic inference is used over all rules to find the most probable assignment to open predicates. The main differences between PSL and DRaiL are - (a) In DRaiL, each rule weight is learned using a neural network, which can take arbitrary input representations, while in PSL a single weight is learned for each rule, and expressive classifiers can only be leveraged as priors; (b) DRaiL defines a shared relational embedding space, by specifying entity and relation specific encoders that are reused across all rules. In both frameworks, rules are transformed into linear inequalities corresponding to their disjunctive form, and MAP inference is defined as a linear program.

In PSL, rules are compiled into a Hinge-Loss Markov random field, defined over continuous variables. Weights can be learned using maximum likelihood estimation, maximum-pseudolikelihood

estimation, or large-margin estimation. In DRaiL, rule weights are learned using neural networks. Parameters can be learned *locally*, by training each neural network independently, or *globally*, by using inference to ensure that the scoring functions for all rules result in a globally consistent decision. To learn global models, DRaiL can also employ maximum likelihood estimation or large-margin estimation. Details regarding both frameworks can be found in Appendix C.2.

5 Experimental Evaluation

The goal of our relational learning framework is to identify morality frames in tweets by modeling them jointly, and derive interpretable relations between them and other contextualizing information. In this section, we compare the performance of our model with multiple baselines, and present a detailed error analysis. Then, we collect tweets on one topic (*Abortion*) and one event (*2021 US Capitol Storming*) written by US Congress members and analyze the discussion.[1] We identify the morality frames in these tweets using our best model.

5.1 Experimental Settings

We experiment with PSL and DRaiL for modeling the rules presented in Section 4.1. In DRaiL, each rule r is associated with a neural architecture, which serves as a scoring function to obtain the rule weight w_r. In the case of rules r_1 and r_2, which map tweets and entities to labels, we use a BERT encoder (Devlin et al., 2019) with a classifier on top. We use task-adaptive pretraining for BERT (Gururangan et al., 2020), and fine-tune it on a large number of unlabeled tweets. In the case of rules r_3 and r_4, that incorporate ideology and topic information, we learn topic and ideology embeddings with one-layer feed-forward nets over their one-hot representations. Then, we concatenate the output of BERT with the topic and ideology embeddings before passing everything through a classifier. On the other hand, PSL directly learns a single weight for each rule. Given that our rules are defined over complex inputs (tweets), we use the output of the locally trained neural nets as priors for PSL, by introducing additional rules of the form $\texttt{Prior}(\texttt{t}, \texttt{x}) \Rightarrow \texttt{Label}(\texttt{t}, \texttt{x})$. This approach has been successfully used in previous work dealing with textual inputs (Sridhar et al., 2015). Note that while PSL can only leverage these classifiers as priors, DRaiL continues to update the parameters of the neural nets during learning.

We model constraint c_1, aligning the MF and role predictions, and c_3, aligning role polarity, as unweighted hard constraints in both frameworks. For constraint c_2, we learn a weight to encourage different entities in a tweet to have different roles. PSL learns a weight directly over this rule, while in DRaiL we use a feed-forward net over the one-hot vector of the relevant MF. We compare our relational models with the following baselines.

Lexicon Matching: Direct keyword matching using the MF Dictionary (MFD) (Graham et al., 2009) and a PMI-based lexicon extracted from the dataset by Johnson and Goldwasser (2018).

Sequence Tagging: We set the MF role prediction task as a sequence tagging problem, and map each entity in a tweet to a role label. We use a BiLSTM-CRF (Huang et al., 2015) over the full tweet, and use the last time-step in each entity span as its emission probability.

End-to-end Classifiers: We map the text and entities, and other contextualizing features (e.g. topic), to a single label. We compare BERT-base and task adaptive pretraining (BERT-tapt) by using a whole-word-masking objective over the large set of unlabeled political tweets.

Multi-task: We define a single BERT encoder, and a single ideology and topic embedding that is shared across the two tasks. Task-specific classifiers are used on top of these representations. Then, the loss functions are added as $L = \lambda_1 L_{\texttt{MF}} + \lambda_2 L_{\texttt{Role}}$. We set $\lambda_1 = \lambda_2 = 1$.

We perform 3-fold cross validation over the dataset introduced in Section 3, and show results for MF and role prediction in Table 7. First, we observe that leveraging unlabeled data for task-adaptive pretraining improves performance. Then, we find that relational models that use probabilistic inference outperform all of the other baselines for both tasks. Further, we find that modeling rules using neural nets in DRaiL, and learning their parameters with global learning, performs better than using them as priors and learning a single weight in PSL. We also include results by fixing the gold labels for the MF prediction, and refer to this as a *skyline*. Unsurprisingly, having perfect MF information improves results for roles considerably. In this case, the candidates for each entity are reduced from 16 possible assignments to 3 or 4, which results in a much easier task. Details regarding all baselines,

[1] Collected from https://github.com/alexlitel/congresstweets

Group	Model	Macro		Weighted	
		Role	MF	Role	MF
Lexicon Matching	MF Dictionary	-	30.37	-	37.32
	PMI Lexicon	-	36.44	-	35.94
	MFD + PMI	-	39.78	-	42.12
Seq-Tagging	BiLSTM-CRF	35.18	-	45.91	-
End-to-end Classifiers	BiLSTM	39.75	58.61	45.61	59.90
	BERT-base	49.32	59.99	57.37	62.17
	BERT-tapt	54.73	66.44	62.18	68.29
	+ Ideo + Issue	54.81	66.13	62.83	68.34
Multi-task	BERT-base	44.37	61.63	57.74	67.71
	BERT-tapt	52.08	63.46	61.96	69.20
	+ Ideo + Issue	52.11	63.44	63.61	68.61
Relational Learning	PSL	56.51	68.98	64.02	71.85
	DRaiL Local	58.07	71.20	64.38	73.85
	DRaiL Global	**59.23**	**72.34**	**64.98**	**74.39**
Skyline	DRaiL Global (Fixed MF)	79.35	-	84.52	-

Table 7: MF and MF role classification F1 Scores.

Models	Weighted F1		# of Errors		
	Role	MF	Polarity Swap (E1)	Mixed MFs (E2)	Same Role (E3)
BERT-tapt	62.18	68.29	274	844	102
All rules	63.93	69.23	260	807	93
$+c_1$	65.11	74.44	254	**732**	130
$+c_2$	63.94	69.37	260	790	101
$+c_3$	64.13	69.31	245	797	**89**
$+c_1 + c_2$	65.04	74.43	254	733	149
$+c_1 + c_3$	**65.32**	**74.53**	249	733	126
$+c_2 + c_3$	63.99	69.22	**244**	791	96
+ All constr	64.98	74.39	248	736	138

Table 8: Ablation Study and Error Analysis.

hyper-parameters, task-adaptive pretraining, and results per class can be found in Appendix D. Code and dataset can be found at https://github.com/ShamikRoy/Moral-Role-Prediction.

5.2 Ablation Study and Error Analysis

We perform an ablation study, evaluating different versions of our model by adding and removing constraints and analyzing corresponding errors. To study the effect of different rules and constraints on role prediction, we define three types of errors:

(E1) Polarity Swap: when the role of an entity with one polarity (positive/negative) is identified as one role of the opposite polarity.

(E2) Mixed MFs: when different entities of the same tweet are identified with roles from a MF other than the gold label of the tweet.

(E3) Same Roles: all of the entities in a tweet are identified to have the same role when the gold labels are different.

The analysis is shown in Table 8. First, we see that constraint c_1, aligning the two decisions, does most of the heavy lifting and reduces error (E2) in all cases. Enforcing consistent polarities with c_3 further improves performance and reduces error (E1), for which it is designed for. c_3 also reduces error (E3) in some models. Encouraging entities to have different roles with c_2 does not improve the overall performance, but it helps to reduce error (E3) when combined with c_3. We use a soft version of c_2, so it is not strictly enforced. We find that roles with negative sentiments are easier for the model to identify (Appendix D.4). Note that every MF has only one role with negative sentiment, and the model does not swap role labels with different sentiments frequently (E1). Therefore, determining the correct positive role is more challenging.

Topic	Tweets	NPs	Ents	Final Tweets	Final (Tw, Ent)
Abortion	18.5k	156	25	9,676	28,054
US Capitol	22.2k	100	25	7,188	14,299

Table 9: Summary of new dataset from US politicians.

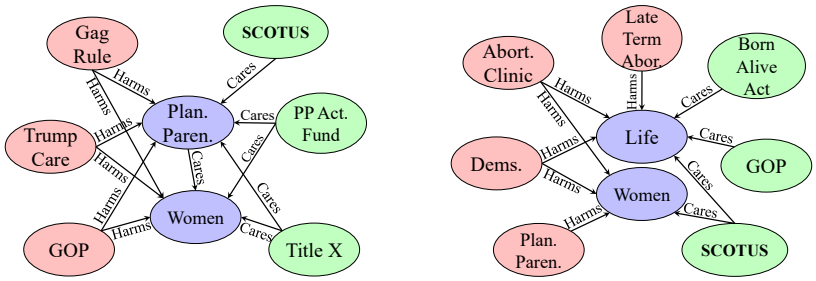
Abortion			US Capitol		
Entity	Freq.	Prec.	Entity	Freq.	Prec.
Women	91	0.89	Trump	2	0.50
P. Parenthood	60	0.75	Congress	58	0.58
Life	31	0.50	Capitol	1	0.57

Table 10: Precision of moral role prediction of entities in new data vs. entity frequency in training data.

5.3 Predicting Morality Frames on New Data

To analyze the political discussion using the moral sentiment towards entities, we collected more tweets from US politicians on the topic of Abortion and around the storming of the US Capitol on Jan. 6, 2021. The Abortion tweets are from 2017 to Feb. 2021. For the US Capitol incident, we collected tweets 7 days before and after the event, with the goal of studying any change in sentiment towards entities. We took noun phrases occurring at least 50 times, manually filtered out non-entities, and grouped different mentions of the same entity (Appendix D.6). We collected tweets that mentioned these entities. Statistics for the resulting data can be found in Table 9. We re-trained our model using all of our labeled data, and predicted the morality frames for each tweet in the new dataset.

We performed human evaluation on the predictions for this new data by randomly sampling 50 tweets from each issue. This resulted in 91 and 76 (tweet, entity) pairs for Abortion and US Capitol,



(a) Entity relationship in left

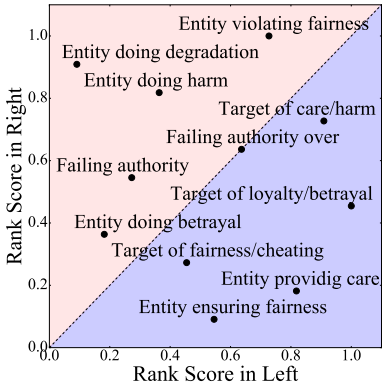
(b) Entity relationship in right

Figure 1: Entity-relation graphs for moral foundation **Care/Harm**. Here, blue, red and green spheres represent target, harming and caring entities, respectively. A directed edge represents relationship between two entities.

respectively. This procedure resulted in an accuracy of MF prediction of 88% for each issue, and a role prediction accuracy of 75% for Abortion, and 60.44% for the US Capitol incident. We found that entities that appear less in the training data have low precision for the role prediction (See Table 10). Note that the US Capitol event was not observed during training, which makes it more challenging. For Abortion, we observed that Democrats mention the entity *Women* most, and 84% of the time the predicted MF role is target of care/harm or fairness/cheating, and it is never assigned a negative role (possibly because of constraint c_3). For Republicans, we observed the same pattern for the entity *Life* (Stats in Appendix D.8). However, in a few cases (2.4%) *Life* is predicted as the entity ensuring fairness/purity/care, justified authority or being loyal. While these roles carry a positive sentiment, they are intuitively wrong predictions for *Life*. We found out that for 34.21% of such cases, there are multiple mentions of *Life* in the same tweet. Given that constraint c_2 encourages different roles for different entities in a tweet, this can be the source of this error. Examples for these cases can be found in Appendix D.9.

	Most Frequent Entities	Most Associated Moral Roles
In Left	Woman	Target of fairness/cheating
	Reproduction Right	Target of fairness/cheating
	Planned Parenthood	Target of loyalty/betrayal
	Reproductive Care	Target of fairness/cheating
	SCOTUS	Entity ensuring fairness
In Right	Life	Target of purity/degradation
	Planned Parenthood	Entity doing cheating
	Democrats	Failing authority
	Born Alive	Target of purity/degradation
	Woman	Target of care/harm

(a) Most frequent entities & most associated moral roles.



(b) Sentiment towards **Planned Parenthood**. Normalized rank scores of MF roles based on usage, are plotted in (x, y)-axes. We discarded roles used <10 times.

Figure 2: Polarization in sentiment towards entities.

6 Analyzing Political Discussions

In this section, we first characterize the political discussion on *Abortion* using the predicted morality frames. Then, we analyze how an event impacts the moral sentiment towards entities by looking at the usage of MF roles before and after the *2021 US Capitol Storming* for the different parties.

6.1 Characterizing Discussion on Abortion

Morality Frame Usage: We found out that the left uses Fairness/Cheating the most, while the right uses Purity/Degradation. Care/Harm is the second most frequent for both parties (Appendix E.1). To analyze MF role usage, we list the most frequent entities and their most frequent moral roles in Figure 2a. The left portrays entities related to *Reproduction Freedom* as the target of Fairness/Cheating. While on the right, the top target of Purity/Degradation is *Life*. Both of them use *Planned Parenthood* frequently, but their sentiment towards it differs. To further examine this, we plot *Planned Parenthood*'s polarity graph in Figure 2b. It shows that parties express opposite sentiments towards *Planned Parenthood*. These findings are consistent with known stances of democrats and republicans on this topic.

Entity-Relation Graph: We examine how the

Entity Types	Top Entities in Left		Top Entities in Right	
	Pre-Event	Post-Event	Pre-Event	Post-Event
Target of Care	Citizens, Democracy, America	Capitol, Democracy, Police	America, Citizens	Capitol, America, Sicknick
Causing Harm	Trump, Violence	Trump, Violence, Domest. terror.	-	Violence, Trump
Provide Care	Congress, Biden, Democrats	Congress, Biden, Amendment	Congress, Trump	Police; Congress
Justified Authority	Congress, Pelosi, Democrats	Congress, Amendment, Pence	Congress	Congress, Pence
Justified Auth. Over	-	Biden, Harris	-	-
Failing Authority	Trump, GOP	Trump, Impeachment, GOP	Democrats, Trump, GOP	Trump, Dems., Impeachment
Failing Auth. Over	Democracy, Biden, McConnell	Democracy, Capitol, Nation	Pelosi, Citizens, America	Nation, Pelosi, Biden

Table 11: Top-3 target entities per role pre and post the US Capitol event on January 6, 2021. Entity roles shown in this table are related to the moral foundations Care/Harm and Authority/Subversion.

political discussion is framed by each party by looking at the sentiments expressed towards different entities, regardless of whether they use the same high level MF. We look at Care/Harm, which is frequently used by both parties, and take the two most used targets by each party. We then take the top three care providing and harming entities used in the same tweet as the target. We assign the most common role for each entity, and represent it in an entity-relation graph in Figure 1a-1b. We can see that both democrats and republicans express care for *Women*, but the caring and harming entities vary highly across parties. For example, the left portrays *Planned Parenthood* as the caring entity, while the right portrays it as the harming entity. This analysis shows that, while there is overlap in the MFs used, the moral roles of entities can highlight the differences between parties in politically polarized discussions at an **aggregate** level.

6.2 Moral Response to US Capitol Storming

To analyze how the moral sentiment towards entities changed after the storming of the US Capitol on January 6, 2021, we look at the sentiment towards entities before and after the event. We found that Authority/Subversion and Care/Harm were the two most used moral foundations after the incident for both parties (Appendix E.1). In Table 11, we present the top three most frequent entities for role types under Care/Harm and Authority/Subversion, before and after the event. Entities appearing less than 15 times are omitted from this analysis. Our model predicted that, after the event, the left justified the authority of *Mike Pence*, and *violence* appeared as a harming entity even before the event occurred. On the right, *Trump* shifted from an entity providing care prior to the event, to a harming entity after the event. We show some relevant tweets and their corresponding predictions in Table 12. The entity-relation graph for each party after the event can can be found in Appendix E.2.

[Ideology-Period] (Predicted MF) Tweet Text
[Left-Pre-Event] (Care/Harm) Zealotry. [Trump]$_{DO-HARM}$ has seeded an anti-government fanaticism among his most fervent followers, threatening systematic [violence]$_{DO-HARM}$ and the future of [American democracy]$_{TARGET-CARE}$.
[Left-Post-Event] (Authority/Subversion) I'm calling on @VP [Pence]$_{JUST-AUTH}$ to invoke the [25th amendment]$_{JUST-AUTH}$ in order to immediately remove President [Trump]$_{FAIL-AUTH}$ from office.
[Right-Pre-Event] (Care/Harm) Great news: [@realDonaldTrump]$_{PROVIDE-CARE}$ just signed the #GLRI Act into law - bipartisan legislation that will help protect & preserve the Great Lakes.
[Right-Post-Event] (Care/Harm) President [Trump's]$_{DO-HARM}$ incendiary rhetoric and false election fraud claims incited his supporters to [violence]$_{DO-HARM}$.

Table 12: Examples of moral roles prediction for entities related to the US Capitol event on January 6, 2021.

7 Summary

In this paper, we present the first study on Moral Foundations Theory at the entity level, by assigning moral roles to entities, and present a novel dataset of political tweets that is annotated for this purpose. We propose a relational model that predicts moral foundations and the moral roles of entities jointly, and show the effectiveness of modeling dependencies and contextualizing information for this task. Finally, we analyze political discussions in the US using these predictions, and show the usefulness of our proposed schema. In the future, we intend to study how morality frames and our relational framework can be applied in other settings, where contextualizing information is not observed.

Acknowledgements

We thank Nikhil Mehta, Rajkumar Pujari, and the anonymous reviewers for their insightful comments. This work was partially supported by an NSF CAREER award IIS-2048001.

8 Ethics Statement

To the best of our knowledge no code of ethics was violated throughout the annotations and experiments done in this paper. We used human annotation for annotating an existing dataset with new labels. We adequately acknowledged the dataset and its various properties are explained thoroughly in the paper. While annotating, we respected the privacy rights of the crowd annotators and we didn't ask any personal details of the anonymous human annotators. They were informed that the task contains potentially sensitive political content. The crowd annotators were fairly compensated by rewards per annotation. We determined what is a fair amount of compensation by taking into consideration the feedback from the annotators and comparing our reward with other annotation tasks on the crowd-sourcing platform.

The dataset presented is comprised of tweets, and for the reviewers, we only submitted a subset of the tweets with text. We will replace the tweet text with only tweet ids when publishing it publicly to respect the privacy policy of Twitter. We did a thorough qualitative and quantitative evaluation of our annotated dataset, presented in the paper. We reported all pre-processing steps, hyper-parameters, other technical details and will release our code and data for reproducibility. Due to space constraints, we moved some of the pre-processing steps, detailed hyper-parameter information, and additional results to the Appendix section. The results reported in this paper support our claims and we believe that they are reproducible. Any qualitative result we report is an outcome from a machine learning model and does not represent the authors' personal views, nor the official stances of the political parties analyzed. As we study text from humans to identify the moral sentiment, to draw conclusions, we rely on a machine learning model which is more interpretable than an end to end deep learning model.

References

Avnika B Amin, Robert A Bednarczyk, Cara E Ray, Kala J Melchiori, Jesse Graham, Jeffrey R Huntsinger, and Saad B Omer. 2017. Association of moral values with vaccine hesitancy. *Nature Human Behaviour*, 1(12):873–880.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452, Berlin, Germany. Association for Computational Linguistics.

Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-loss Markov random fields and probabilistic soft logic. *Journal of Machine Learning Research (JMLR)*.

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482, Denver, Colorado. Association for Computational Linguistics.

Amber Boydstun, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues.

William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Dennis Chong and James N Druckman. 2007. Framing theory. *Annu. Rev. Polit. Sci.*, 10:103–126.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Morteza Dehghani, Kate Johnson, Joe Hoover, Eyal Sagi, Justin Garten, Niki Jitendra Parmar, Stephen Vaisey, Rumen Iliev, and Jesse Graham. 2016. Purity homophily in social networks. *Journal of Experimental Psychology: General*, 145(3):366.

Morteza Dehghani, Kenji Sagae, Sonya Sachdeva, and Jonathan Gratch. 2014. Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "ground zero mosque". *Journal of Information Technology & Politics*, 11(1):1–14.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2970–3005.

Lingjia Deng and Janyce Wiebe. 2015. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 179–189, Lisbon, Portugal. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. *arXiv preprint arXiv:1906.01762*.

Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3730–3736, Portorož, Slovenia. European Language Resources Association (ELRA).

Justin Garten, Reihane Boghrati, Joe Hoover, Kate M Johnson, and Morteza Dehghani. 2016. Morality between the lines: Detecting moral sentiment in text. In *Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes*.

Jesse Graham, Jonathan Haidt, and Brian Nosek. 2009. Graham, Haidt, & Nosek (2009): Liberals and conservatives rely on different sets of moral foundations.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 434–444, Hong Kong, China. Association for Computational Linguistics.

Joe Hoover, Kate Johnson, Reihane Boghrati, Jesse Graham, Morteza Dehghani, and M Brent Donnellan. 2018. Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation. *Collabra: Psychology*, 4(1).

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Kristen Johnson and Dan Goldwasser. 2016. Identifying stance by analyzing political discourse on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 66–75.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality and quantity*, 38:787–800.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring background knowledge to improve moral value prediction. In *2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)*, pages 552–559. IEEE.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2019. Discourse representation structure parsing with recurrent neural networks and the transformer model. In *Proceedings of the IWCS Shared Task on Semantic Parsing*, Gothenburg, Sweden. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

André F. T. Martins, Mário A. T. Figueiredo, Pedro M. Q. Aguiar, Noah A. Smith, and Eric P. Xing. 2015. Ad3: Alternating directions dual decomposition for map inference in graphical models. *Journal of Machine Learning Research*, 16(16):495–545.

Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, 2(6):389–396.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

Maria Leonor Pacheco and Dan Goldwasser. 2021. Modeling content and context with deep relational learning. *Transactions of the Association for Computational Linguistics*, 9:100–119.

Chan Young Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov. 2020. Multilingual contextual affective analysis of lgbt people portrayals in wikipedia. *arXiv preprint arXiv:2010.10820*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine learning*, 62(1-2):107–136.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7698–7716, Online. Association for Computational Linguistics.

Shamik Roy and Dan Goldwasser. 2021. Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 1–13, Online. Association for Computational Linguistics.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 116–125, Beijing, China. Association for Computational Linguistics.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2018. Hierarchical structured model for fine-to-coarse manifesto text analysis. *arXiv preprint arXiv:1805.02823*.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China. Association for Computational Linguistics.

David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. 2015. Structured training for neural network transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 323–333, Beijing, China. Association for Computational Linguistics.

Manuel Widmoser, Maria Leonor Pacheco, Jean Honorio, and Dan Goldwasser. 2021. Randomized deep structured prediction for discourse-level processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1174–1184, Online. Association for Computational Linguistics.

Christopher Wolsko, Hector Ariceaga, and Jesse Seiden. 2016. Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology*, 65:7–19.

Jing Yi Xie, Renato Ferreira Pinto Junior, Graeme Hirst, and Yang Xu. 2019. Text-based inference of moral sentiment change. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4646–4655.

Xiao Zhang, Yong Jiang, Hao Peng, Kewei Tu, and Dan Goldwasser. 2017. Semi-supervised structured prediction with neural CRF autoencoder. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1701–1711, Copenhagen, Denmark. Association for Computational Linguistics.

A Data Collection

Identification and Annotation of Tweets with 'Purity/Degradation': To collect more tweets on Purity/Degradation, we took more examples from the unlabeled segment of the dataset (93K tweets). Then we filtered out 619 tweets from it based on lexicon matching with Moral Foundation Dictionary for Purity/Degradation. Then two of the authors of this paper individually went over the 619 tweets and selected tweets having purity/degradation as the primary moral foundation in them. The two authors had agreement on 95% of the cases. Then we combined the two lists from two authors and in case of a disagreement we resolved it by discussion. In this manner we found 44 tweets on Purity/Degradation. Then we annotate these 44 tweets with Purity/Degradation with 17 Policy Frames present in them in the same manner. Two authors of this paper annotated the 44 tweets for Policy Frames individually. They had an agreement on 47% of the cases about the primary policy frame in a tweet. Most of the time they had a disagreement in the cases where there are more than 1 policy frame present in them. The authors resolved any disagreement by discussion.

Full Dataset Statistics: The statistics of the full dataset can be found in Table 13.

MORALS	# OF TWEETS	IDEOLOGY		TOPIC					
		LEFT	RIGHT	ABO	ACA	GUN	IMM	LGBT	TER
Care/Harm	589	378	211	30	142	221	31	11	154
Fairness/Cheating	264	201	63	42	81	33	22	73	13
Loyalty/Betrayal	231	167	64	15	20	92	28	24	52
Authority/Subversion	471	200	271	33	177	76	99	19	67
Purity/Degradation	44	13	31	21	3	8	6	2	4
TOTAL	1599	959	640	141	423	430	186	129	290

Table 13: Dataset summary.

B Data Annotation for Moral Roles

B.1 Questionnaire asked to the annotators for annotation of entity roles

The questionnaire asked to the annotators for all moral foundations can be found in Table 14.

B.2 Calculation of Partisanship and Most Frequent Entities by Entity Role

To determine the partisanship of the elements - (1) moral foundations, (2) (moral foundation role: entity), we use z-score measure of these elements in the two political ideologies (left, right). We calculate the z-score to evaluate - whether two groups (e.g., left and right) differ significantly on some single characteristic. In our case the characteristics are any element of type (1) or type (2) as described above. A positive z-score means it's left-partisan and negative score means right-partisan.

Most frequent entities per moral role can be found in Table 15.

B.3 Expressivity of bias of Moral Roles

To examine how well moral roles account for political standpoints when compared to moral foundations, we use the moral foundations (MF) and (moral foundation role, entity) (MFR) as one hot encoded features to classify the ideology of the tweet (left/right). The results are shown in Tab. 16. Moral roles classify the ideology reasonably well compared to MF and BoW features, which proves the usefulness of the moral roles for capturing political perspectives.

C Modeling

C.1 Polarity of Moral Roles

Moral Roles with positive polarity: Target of care/harm, Entity providing care, Target of fairness/cheating, Entity ensuring fairness, Target of loyalty/betrayal, Entity being loyal, Justified authority, Justified authority over, Failing authority over, Target of purity/degradation, Entity preserving purity.

MORAL	ENTITY TYPE	QUESTION ASKED TO THE ANNOTATORS
Care/Harm	Target of care/harm	Which entity needs CARE, or is being HARMED?
	Entity causing harm	Which entity is causing the HARM?
	Entity providing care	Which entity is offering/providing the CARE?
Fairness/Cheating	N/A (additional question)	Fairness or cheating on what?
	Target of fairness/cheating	Fairness for whom or who is being cheated?
	Entity ensuring fairness	Who or What is ensuring fairness or in charge of ensuring fairness?
	Entity doing cheating	Who or What is cheating or violating the fairness?
Loyalty/Betrayal	N/A (additional question)	What are the phrases invoking LOYALTY?
	N/A (additional question)	What are the phrases invoking BETRAYAL?
	Target of loyalty/betrayal	LOYALTY or BETRAYAL to whom or what?
	Entity being loyal	Who or what is expressing LOYALTY?
	Entity doing betrayal	Who or what is doing BETRAYAL?
Authority/Subversion	N/A (additional question)	LEADERSHIP or AUTHORITY on what issue or activity?
	Justified authority	Which LEADERSHIP or AUTHORITY is obeyed/praised/justified?
	Justified authority over	If the LEADERSHIP or AUTHORITY is obeyed/praised/justified, then praised/obeyed by whom or justified over whom?
	Failing authority	Which LEADERSHIP or AUTHORITY is disobeyed or failing or criticized?
	Failing authority over	If the LEADERSHIP or AUTHORITY is disobeyed or failing or criticized, then failing to lead whom or disobeyed/criticized by whom?
Purity/Degradation	Target of purity/degradation	What or who is SACRED, or subject to degradation?
	Entity preserving purity	Who is ensuring or preserving the sanctity?
	Entity causing degradation	Who is violating the sanctity or who is doing degradation or who is the target of disgust?

Table 14: Questionnaire asked to Amazon Mechanical Turk annotators for annotation of entities.

Moral Roles with negative polarity: Entity causing harm, Entity doing cheating, Entity doing betrayal, Failing authority, Entity causing degradation.

C.2 Relational Learning Frameworks

C.2.1 Probabilistic Soft Logic

PSL models are specified using weighted horn clauses, which are compiled into a Hinge-Loss Markov Random Field, a class of undirected probabilistic graphical model. In HL-MRFs, a probability distribution is defined over continuous values in the range of [0, 1], and dependencies among them are modeled using linear and quadratic hinge functions. This way, they define a probability density function:

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \frac{1}{Z} \exp\left(-\sum_{r=1}^{M} w_r \psi_r(\boldsymbol{X}, \boldsymbol{Y})\right) \qquad (1)$$

where w_r is the rule weight, Z is a normalization constant and $\psi_r(\boldsymbol{Y}, \boldsymbol{X}) = \max\{l_r(\boldsymbol{X}, \boldsymbol{Y}), 0\}^{\rho_r}$ is the hinge-loss potential corresponding to the instantiation of rule r, represented by a linear function l_r of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and an optional exponent $\rho_r \in \{1, 2\}$. Inference in PSL is performed by finding a MAP estimate of the random variables $\boldsymbol{Y}$ given evidence $\boldsymbol{X}$, this is done by maximizing the density function in Eq. 1 as $\arg\max_{\boldsymbol{Y}} P(\boldsymbol{Y}|\boldsymbol{X})$. To solve this, they use Alternating Direction Method of Multipliers (ADMM), an efficient convex optimization procedure.

Weights can be learned through maximum likelihood estimation by using the structured perceptron algorithm. The partial derivative of the log of the likelihood function in Eq. 1 above with respect to a parameter $\boldsymbol{W}_r$ is:

$$\frac{\partial \log P(\boldsymbol{Y}|\boldsymbol{X})}{\partial \boldsymbol{W}_r} = \mathbb{E}_{\boldsymbol{W}}[\psi_r(\boldsymbol{X}, \boldsymbol{Y})] - \psi_r(\boldsymbol{X}, \boldsymbol{Y}) \qquad (2)$$

where $\mathbb{E}_{\boldsymbol{W}}$ is the expectation under the distribution defined by $\boldsymbol{W}$. Given that computing this expectation is intractable, they approximate it by taking the values in the MAP state. This approximation makes this learning approach a structured variant of the voted perceptron. Note that alternative estimations are also supported. More details can be found in the original paper (Bach et al., 2017).

C.2.2 DRaiL

Rules in DRaiL can be *weighted* (i.e. classifiers, soft constraints) or *unweighted* (i.e. hard constraints). The collection of all rules represents the global decision. Rules are transformed into linear inequalities, corresponding to their disjunctive form, and MAP inference is then defined as an

MORAL	ENTITY TYPE	MOST FREQUENT ENTITY IN LEFT	MOST FREQUENT ENTITY IN RIGHT
Care/Harm	Target of care/harm	20 million Americans; our families; woman; innocent people; #domesticviolence victims	law-abiding Americans; victims and their families; small businesses; patients; Paris
	Entity causing harm	gun show loopholes; gun violence; terrorist attack; mass-shootings; suspected terrorists	Radical Islamic terrorists; #Obamacare mandates; Brussels attacks; #ISIS; ISIL-Inspired Attacks
	Entity providing care	gun show loophole bills; Affordable Care Act; #ImmigrationReform; Democrats; commonsense gun legislation	@RepHalRogers: Bill; @HouseGOP; Senate; @WaysandMeansGOP; HR 240
Fairness/Cheating	Target of fairness/cheating	woman, #LGBT community; all Americans; #FightForFamilies; other vulnerable people	the American people; small businesses; people; religious minorities in Syria and Iraq
	Entity ensuring fairness	#SCOTUS decision; congress; bill to expand access; the DREAM Act; Equality Act	Senate; House; @RepHalRogers: Bill; House GOP; Supreme Courtś #HobbyLobby ruling
	Entity doing cheating	anti-#LGBT laws; employer; HB 2; #HobbyLobby decision; Political attacks	#Obamacare legislation; fake ISIS passports; Planned Parenthood; the Pakistani Gov; enforcement loopholes
Loyalty/Betrayal	Target of loyalty/betrayal	#LGBT communities; gun safety measures; victims of #Orlando; women men and families; #StandwithPP	Paris terror attacks; senators; Israel; The American people; Syrian and Iraqi refugees
	Entity being loyal	@SenWarren; @RepAdams; My colleagues; @SenateDems; House Democrats	@SenatorIsakson; @RepHalRogers
	Entity doing betrayal	@HouseGOP extremist Members!; terrorists; The community of nations; @NRA	House
Authority/Subversion	Justified authority	POTUS; SCOTUS; President Obamaś; Senate; @HouseDems	@HouseGOP; #Senate; #SCOTUS; Congress; Republicans
	Justified authority over	Americans; 180 House Dems; nation; people	@SenateMajLdr; @RepHalRogers; #American; @SenateMajLdr; Inhofe
	Failing authority	#HouseGOP; Congress; Republicans; SCOTUS; @SpeakerRyan	President Obama; POTUS; #Obamacare; @SCOTUS; @SecBurwellś
	Failing authority over	Americans; @repjohnlewis; family; @SenFeinsteinś; women; Sen Dems	Americans; @SenateMajLdr; @HouseAppropsGOP @RepHalRogers; @SenateMajLdr McConnell; @SpeakerRyan
Purity/Degradation	Target of purity/degradation	immigration; women	fetal body parts; lives of the unborn; baby girls
	Entity preserving purity	N/A (no ngrams found that occurs more than 2 times.)	@SenDanCoats; #MarchforLife
	Entity causing degradation	Donald Trump; Charleston church killings	Planned Parenthood; abortion providers; Radical Islamic terrorists

Table 15: Most frequent entities by ideology and by moral role type. For each ideology the most frequent list was generated by taking the most common stemmed ngrams (n = 1 to 5) in the identified entities by the annotators. One representative entity from each ngram group is presented in this table. Inclusive ngrams were merged together. For example: 'law abiding citizen' is merged with 'law abiding'.

One-hot Encoded Features	# of Features	Macro F1
Moral Foundation (MF)	5	0.62
Moral Roles (MFR)	2021	0.77
MF+MFR	2026	0.79
Bag of Words (BoW)	2478	0.85
BoW+MF	2483	0.85
BoW+MFR	4499	0.87
BoW+MF+MFR	4504	0.86

Table 16: Predicting ideology of tweet using Logistic Regression (3-fold CV).

integer linear program:

$$\begin{aligned} \arg\max_{\boldsymbol{y}\in\{0,1\}^n} P(\boldsymbol{y}|\boldsymbol{x}) \equiv \arg\max_{\boldsymbol{y}\in\{0,1\}^n} & \sum_{\psi_{r,t}\in\Psi} w_r\ \psi_r(\boldsymbol{x_r}, \boldsymbol{y_r}) \\ s.t.\ & c(\boldsymbol{x_c}, \boldsymbol{y_c}) \leq 0;\ \ \forall c \in C \end{aligned} \tag{3}$$

Where each rule grounding r, generated from template t, with input features $\boldsymbol{x_r}$ and predicted variables $\boldsymbol{y_r}$ defines the potential $\psi_r(\boldsymbol{x_r}, \boldsymbol{y_r})$, added to the linear program with a weight w_r. DRaiL implements both exact and approximate inference to solve the MAP problem, in the latter case, the AD^3 algorithm is used (Martins et al., 2015).

In DRaiL, weights w_r are learned using neural networks defined over parameter set $\boldsymbol{\theta}$. Parameters can be learned *locally*, by training each rule independently, or *globally*, by using inference to ensure that the scoring functions for all rules result in a globally consistent decision. To train global models using large-margin estimation, DRaiL uses the structured hinge loss:

$$\max_{\hat{\mathbf{y}}\in Y}(\Delta(\hat{\mathbf{y}}, \mathbf{y}) + \sum_{\psi_r\in\Psi} \Phi_t(\boldsymbol{x_r}, \hat{\boldsymbol{y}}_{\boldsymbol{r}}; \theta^t)) - \sum_{\psi_r\in\Psi} \Phi_t(\boldsymbol{x_r}, \boldsymbol{y_r}; \theta^t)$$

Where Φ_t represents the neural network associated with rule template t, and parameter set θ^t. Here, $\mathbf{y}$ corresponds to the gold assignments, and $\hat{\mathbf{y}}$ corresponds to the prediction resulting from the MAP inference defined in Eq. 3. Note that alternative estimations are also supported. More details can be found in the original paper (Pacheco and Goldwasser, 2021).

D Experimental Evaluation

D.1 Task-Adaptive Pretraining

We do task-adaptive pretraining for BERT (Gururangan et al., 2020), and fine-tune it on a large number of unlabeled tweets[2]. To select unlabeled

[2]Collected From: https://github.com/alexlitel/congresstweets

tweets, we build a topic-specific lexicon of n-grams (n$\leq$ 5) from our training dataset based on Point-wise Mutual Information (PMI) scores (Church and Hanks, 1990). Namely, for an ngram w we calculate the point-wise mutual information (PMI) with label l (e.g. topic), $I(w, l)$ using the following formula.

$$I(w, l) = \log \frac{P(w|l)}{P(w)}$$

Where $P(w|l)$ is computed by taking all tweets with label l and computing $\frac{count(w)}{count(allwords)}$. Similarly, $P(w)$ is computed by counting ngram w over the set of tweets with any label. To construct the lexicon, we rank ngrams for each label based on their PMI scores.

We explore three pretraining objectives, described below. In all cases, models were initialized using BERT (Devlin et al., 2019).

Masked Language Modeling: We randomly mask some of the tokens from the input, and predict the original vocabulary id of the masked word based on its context (Devlin et al., 2019).

Whole Word Masking: Instead of masking randomly selected tokens, which may be sub-segments of words, we mask randomly selected words.

Moral Foundations Dictionary: We create a lexicon for each Moral Foundation from the dataset by Johnson and Goldwasser (2018) using the same PMI formula described above. We use the normalized PMI scores as a weight for each unigram, and assign a weight of 1 to unigrams in the Moral Foundation Dictionary (MFD)(Graham et al., 2009). We score a tweet by summing the scores of words matching the lexicon. We take the highest scoring moral foundation for each tweet, and fine-tune a moral foundation classifier using this weakly annotated data.

We evaluate these objectives by performing the pre-training stage on the unlabeled data, and fine-tuning the encoder for our base task of leveraging only text to predict moral foundations and entity roles. Results can be seen in Tab. 17.

Objective	Macro		Weighted	
	Role	MF	Role	MF
BERT-base-uncased	49.32	59.99	57.37	62.17
Masked LM	53.49	63.90	61.51	67.45
Whole Word Masking	**54.73**	**66.44**	**62.18**	**68.29**
MF Dictionary	47.92	63.70	54.93	65.81

Table 17: Task-Adaptive Pretraining (F1 Scores)

D.2 Details About the Baselines

Lexicon Matching: We label a tweet with the moral foundation with maximum score based on lexicon matching. We use the Moral Foundation lexicons created in Appendix D.1. If there is no lexicon matching for a tweet, we assign a moral foundation label to it randomly. We experiment with combining and not combining the Moral Foundations Dictionary (MFD) (Graham et al., 2009).

Sequence Tagging: We use a bidirectional LSTM with a CRF layer on top for tagging each entity a tweet with a moral role label. We run two LSTMs in forward and reverse direction of a tweet and concatenate the hidden states (50d) of two directions at each time step to get an embedding (100d) of the token. Given that entity spans are known, we use the last token in each entity as the entity embedding. This embedding is then used for the CRF layer.

End-to-end Classifiers: For the classification of moral foundations using BiLSTM, we run two opposite directional LSTMs over the GloVe word embeddings (Pennington et al., 2014) of all tokens of a tweet, concatenate the hidden states (150d) of both LSTMs to get the embedding of a token (300d), then average the embeddings of all tokens to get a final embedding of a tweet. Then we use this embedding to classify the tweet in the moral foundation classes using a fully connected layer that maps the embedding to a moral foundation class. For moral foundation role classification using BiLSTM, we repeat the same process for an entity text to get its representation using BiLSTM. Then we concatenate the tweet representation and the entity representation and pass it through a hidden layer to get a representation of size 300. Then we use this representation for classification of moral foundation roles using a fully connected layer that maps the representation to the moral foundation role classes. For BERT based models, we use a classifier on top of the `[CLS]` representation. For role classification, we pass an input of the form `[CLS] [tweet] [SEP] [entity]`. We use the default parameters of the BERT-base-uncased huggingface implementation.

Multitasking Based: We define a single BERT encoder, and a single ideology and topic embedding that is shared across the two tasks. The three representations are concatenated and task-specific

Task	Param	Search Space	Selected Value
Local (Base)	Learning Rate	5e-5, 2e-5, 1e-5, 1e-6	2e-5
	Batch size	64, 32	32
	Patience	3, 5, 10	10
	Optimizer	SGD,Adam,AdamW	AdamW
	Hidden Units	-	100
	Non-linearity	-	ReLU
Local (Soft Constr.)	Learning Rate	1e-3, 5e-3, 5e-2, 1e-2	5e-3
	Batch size	64, 32	32
	Patience	5, 10, 20	20
	Optimizer	SGD,Adam,AdamW	AdamW
	Hidden Units	-	100
	Non-linearity	-	ReLU
DRail Global	Learning Rate	5e-5, 2e-5, 1e-5, 1e-6	1e-6
	Batch size	-	Full instance
	Patience	3, 5, 10	10
	Optimizer	SGD,Adam,AdamW	AdamW
	Hidden Units	-	100
	Non-linearity	-	ReLU

Group	Model	Macro		Weighted	
		Role	MF	Role	MF
Simple	BERT	52.37	60.38	63.26	67.57
	+ Ideo + Issue	52.52	60.31	64.02	66.58
	Combined	53.34	59.84	64.65	67.29
Struct.	DRaiL Local	51.71	64.02	63.99	71.37
	DRail Global	**53.23**	**65.46**	**65.50**	**72.39**
Skyline	DRaiL Global (Fixed MF)	76.85	-	86.27	-

Table 18: Hyper-parameter tuning (top) and validation set performance on the best model that combines all rules and constraints in DRaiL (bottom).

classifiers are used on top of them. Then, the loss functions are added as $L = \lambda_1 L_{\texttt{MF}} + \lambda_2 L_{\texttt{Role}}$. We set $\lambda_1 = \lambda_2 = 1$. For topic and ideology embeddings, we use feed-forward computations with 100 hidden layers and ReLU activations. For BERT we use the same configuration as the end-to-end classifiers.

D.3 Hyper-parameter Tuning and Validation Set Performance

For the underlying BERT, we use the default parameters of the hugging face implementation[3]. Other parameters can be observed in Table 18 (top). The bottom part of Table 18 shows the validation performance during the learning of the best performing model.

D.4 Results per Class

The per class classification results can be found in the Table 19.

[3] https://github.com/huggingface/transformers

MF	Pre.	Rec.	F1	Sup.
Auth/Subv	84.64	78.31	81.35	415
Care/Harm	72.10	84.30	77.72	567
Fair/Cheat	70.71	56.91	63.06	246
Loyal/Betray	66.18	63.68	64.90	212
Purity/Degrad	84.85	66.67	74.67	42
Macro Avg.	**75.69**	**69.97**	**72.34**	1482
Weight Avg.	**74.89**	**74.63**	**74.39**	

MF	Role	Pre.	Rec.	F1	F1-Sky	Sup.
Auth/ Subv	Justified	56.43	50.64	53.38	67.09	156
	Justified Over	47.69	46.27	46.97	64.52	67
	Fail	71.96	71.43	71.69	80.90	273
	Fail Over	67.76	65.91	66.82	80.56	220
Care/ Harm	Target	67.42	78.01	72.33	92.72	382
	Cause Harm	82.41	8159	82.00	92.70	402
	Provide Care	57.11	74.23	64.56	91.06	357
Fair/ Cheat	Target	68.18	59.66	63.64	92.01	176
	Ensure Fair	67.22	54.02	59.90	91.65	224
	Do Cheat	64.62	43.75	52.17	83.87	96
Loyal/ Betray	Target	55.47	54.87	55.17	77.96	277
	Be Loyal	59.28	56.25	57.73	74.64	176
	Do Betray	17.65	16.22	16.90	32.50	37
Purity/ Degrad	Target	60.00	56.76	58.33	81.01	37
	Preserve Purity	86.67	46.43	60.47	83.02	28
	Cause Degrad	77.78	56.76	65.62	83.33	37
	Macro Avg.	**62.98**	**57.05**	**59.23**	79.35	2945
	Weight Avg.	**65.56**	**65.23**	**64.98**	84.52	

Table 19: Moral foundation classification results per class (top) and role classification results per class (bottom).

D.5 Run-time Analysis

All experiments were run on a 4 core Intel(R) Core(TM) i5-7400 CPU @ 3.00GHz machine with 64GB RAM and an NVIDIA GeForce GTX 1080 Ti 11GB GDDR5X GPU. Runtimes for our models can be found in Table 20

Task	sec p/Epoch	epochs p/It	sec p/It
r_1 local	9.588	18	177.037
r_2 local	4.002	15	64.106
r_3 local	9.762	27	269.350
r_4 local	4.286	17	74.541
c_2 local	0.138	68	144.631
Global learn	49.259	25	1268.615
Global predict	-	-	7.725

Table 20: Average runtimes for 3 fold cross-validation

D.6 Entity Groups

D.6.1 Entity Groups For Abortion

Brett Kavanaugh: brett kavanaugh, kavanaugh, stop kavanaugh
Roe v Wade: roe v wade, commit roe, protect roe
Planned Parenthood: plan parenthood, stand pp, pp, ppfa, ppact
Affordable Care Act: aca, affordable health care
Title X: title x, family planning, protect x
Gag Rule: gag rule, global gag rule, domestic gag rule

Democrats: democrat, dem, house democrat
Republicans: republican, house gop, senate gop, gop, gop leader
Trump care: trump care
Woman: woman
Reproductive Right: reproductive right, woman reproductive right, reproductive freedom, reproductive justice, woman reproductive freedom
Reproductive Health: reproductive health, woman reproductive health, reproductive health care, reproductive care, reproductive healthcare, reproductive health service, comprehensive reproductive health care, abortion care
SCOTUS: scotus, save scotus, supreme court, supreme court justice, supreme court decision
Life: human life, innocent life, stand life, unborn child, unborn child protection, unborn baby, unborn, baby
NRLC: nrlc
NARAL: naral
Born Alive: bear alive abortion, bear alive
Late Term Abortion: late term abortion
Late Term Abortion Ban: week abortion ban
Born Alive Act: bear alive act
Abortion Provider: abortion provider, abortion clinic, abortion industry
Hyde Amendment: hyde, hyde amendment, bold end hyde
Healthcare Decision: health care decision, healthcare decision
Medicaid: medicaid
Medicare: medicare

D.6.2 Entity Groups for the 2021 US Capitol Hill Storming Event

Congress: congress, th congress
POTUS: potus, president
Donald Trump: trump, donald trump, real donald trump, president real donald trump
America: america
American People: american people
Democracy: american democracy, democracy
Joe Biden: joe biden, biden, president elect
Amendment: amendment, th amendment
Brian Sicknick: brian sicknick, sicknick
Nancy Pelosi: pelosi, speaker pelosi, nancy pelosi
Jamie Raskin: raskin
Capitol: capitol, capitol building, capitol hill, nation capitol
Impeachment: impeachment, impeach president
Kamala Harris: kamala harris, vice president elect
Capitol Police: capitol police, police officer, law enforcement, law enforcement officer
Mike Pence: pence, vp pence, mike pence
Mitch McConnell: mitch mcconnell, mcconnell
GOP: house gop, gop leader, gop, republican
Domestic Terrorism: domestic terrorist, domestic terrorism
Nation: nation
National Security: national security, national guard
Democrats: dem, democrat, house democrat
Violence: violence, violent insurrection, violent attack, violent mob
White Supremacist: white supremacist
Fair Election: fair election

D.7 Human Evaluation on Test Data

Model Prediction Validation We trained our model with all of our labeled data and used it to predict the moral foundations and entity roles of (tweet, entity) pairs in the new set. The validation set (randomly selected from train set) weighted F1 scores were 72.20% and 64.59% for moral foundations and roles, respectively. We validate our model's prediction on the unseen dataset using human evaluation. We randomly sampled 50 tweets from each of the two test sets. This resulted in 91 and 76 (tweet, entity) pairs for Abortion and US Capitol, respectively. Note that one tweet may have > 1 entities. Then, we presented the predictions of moral foundations and entity roles to two graduate students and asked them if the prediction is correct or not. We found the Cohen's Kappa (Cohen, 1960) score between the annotators to be 0.50 (moderate agreement) and 0.64 (substantial agreement) in case of the moral foundations and entity roles, respectively. In case of a disagreement, we asked a third grad student to break the tie. The accuracy of the model for moral foundations was 88% for each topic, while for roles it was 75% and 60.44%, for Abortion and US Capitol, respectively.

D.8 Distribution of MF Roles Assigned by the Model to 'Women' and 'Life'

Distribution of MF Roles Assigned by the Model to 'Women' and 'Life' when mentioned by Democrats and Republicans, respectively, are shown in Table 22 and Table 23, respectively.

(Predicted MF) Tweet	Comment
(CARE/HARM) The U.S. Senate is set to vote on commonsense legislation to protect [unborn babies]$_{TARGET-OF-CARE}$ who can feel pain. Retweet if you Stand For [Life]$_{PROVIDING-CARE}$!	MF prediction is 'Care/Harm', possibly because there is a notion of protecting babies. In MF role prediction, the model makes mistake when there are multiple mention of the same entity, possibly because of constraint c_2 but still assigns a positive role to 'Life', possibly because of constraint c_3.
(LOYALTY/BETRAYAL) I will always, always, ALWAYS be proud to Stand 4 [Life]$_{BEING-LOYAL}$. I'm so grateful to @TXRightToLife for their support and pledge to never stop fighting for the [unborn]$_{TARGET-OF-LOYALTY}$. Now, Texas, let's get out and vote to #KeepTexasRed!	MF prediction is correct. In MF role prediction, the model makes mistake when there are multiple mention of the same entity, possibly because of constraint c_2 but still assigns a positive role to 'Life', possibly because of constraint c_3.
(PURITY/DEGRADATION) [Planned Parenthood]$_{VIOLATING-PURITY}$ is suing our state to expand their abortion-on-demand agenda. RT if you stand for [life]$_{PRESERVING-PURITY}$!	MF prediction is wrong. Still a positive role is assigned to 'life' and a negative role is assigned to 'Planned Parenthood', possibly because of constraint c_3.

Table 21: Qualitative error analysis for MF role prediction for the entity 'Life' in Republicans.

Moral Roles	% of time Assigned by the Model
Target of fairness/cheating	0.624
Target of care/harm	0.216
Failing authority over	0.076
Target of loyalty/betrayal	0.075
Target of purity/degradation	0.006
Entity providing care	0.001
Entity being loyal	0.001

Table 22: Distribution of MF roles assigned to 'Women' by the model when mentioned by the 'Democrats'.

Moral Roles	% of time Assigned by the Model
Target of purity/degradation	0.501
Target of care/harm	0.266
Target of loyalty/betrayal	0.151
Target of fairness/cheating	0.038
Failing authority over	0.018
Entity being loyal	0.008
Entity providing care	0.005
Entity preserving purity	0.004
Justified authority	0.004
Entity ensuring fairness	0.003
Justified authority over	0.002

Table 23: Distribution of MF roles assigned to 'Life' by the model when mentioned by the 'Republicans'.

Moral Foundations	%Used in Left	%Used in Right	%Predicted in Total
Care/Harm	0.23	0.24	0.24
Fairness/Cheating	0.52	0.14	0.35
Loyalty/Betrayal	0.09	0.17	0.13
Authority/Subv.	0.13	0.12	0.13
Purity/Degrad.	0.02	0.34	0.17

Morals	%Used in Left		%Used in Right		%Pred. in Total
	Pre	Post	Pre	Post	Pre+Post
Care/Harm	0.33	0.4	0.27	0.42	0.37
Fairness/Cheating	0.05	0.02	0.04	0.03	0.03
Loyalty/Betrayal	0.19	0.15	0.22	0.25	0.19
Authority/Subv.	0.42	0.42	0.46	0.29	0.40
Purity/Degrad.	0.01	0.01	0.02	0.01	0.1

Table 24: Moral foundation usage by ideologies on topic Abortion (top); and pre and post the US Capitol incident on Jan 6, 2021 (bottom). The percentage of time each moral foundation is predicted by the model are shown in the right-most column of each table.

D.9 Qualitative Evaluation of MF Role Prediction for 'Life' in Republicans

Some tweets mentioning 'Life' by the Republicans and the predicted MF and MF roles are shown in Table 21.

E Analysis of Political Discussion

E.1 Moral Foundation Usage

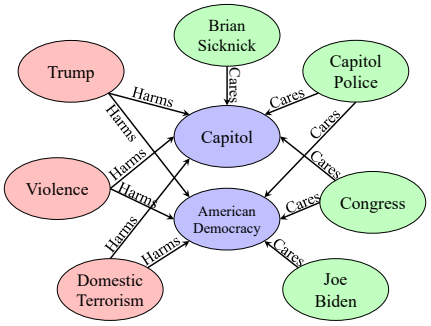
Wikipidia link to the US Capitol incident: Link. The distribution of the usage of different moral foundations on the topics Abortion and US Capitol event can be found in Table 24 (top) and Table 24 (bottom), respectively.

E.2 Most Targeted Entities and Entity Relationship Graphs After the Event US Capitol Storming (2021)

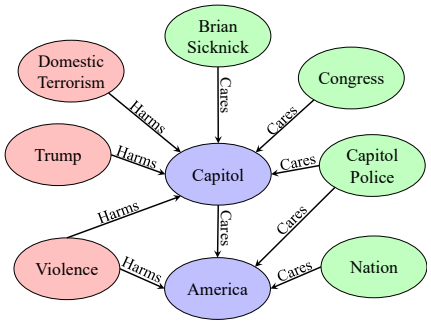
The most targeted entities and entity relationship graphs after the US Capitol Storming (2021) are shown in Figure 3 and 4, respectively.

	Most Frequent Entities	Most Associated Moral Roles
In Left	Trump	Failing authority
	Capitol	Target of care/harm
	Democracy	Target of care/harm
	Congress	Justified authority
	Police	Entity providing care
In Right	Police	Target of loyalty/betrayal
	Capitol	Target of care/harm
	Trump	Failing authority
	Violence	Entity doing harm
	Nation	Target of care/harm

Figure 3: Most frequent entities & most associated MF roles after the event US Capitol Storming (2021).



(a) Entity relation in left



(b) Entity relation in right

Figure 4: Entity relationship graphs for **Care/Harm** after the US Capitol Storming (2021).